2012年6月15日　統計数理研究所　オープンハウス

財務データに対する欠損値補間方法について

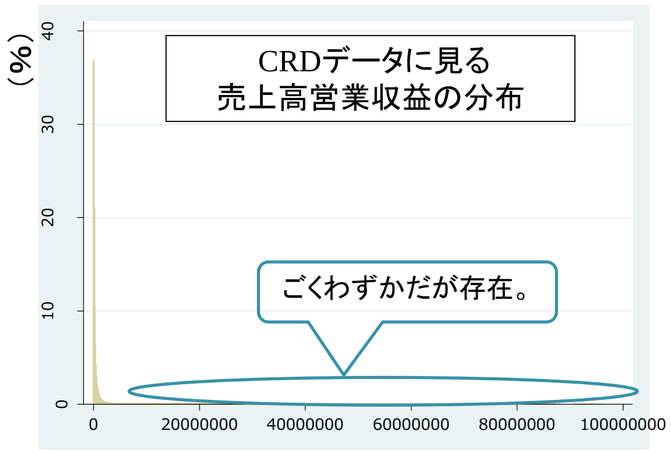
高橋　淳一　　総合研究大学院大学　統計科学専攻　博士後期課程4年

【はじめに】

1500万件以上の中小企業の財務データ（決算書情報）を蓄積しているCRD(Credit Risk Database)のデータではあるが、これまではあまり学術研究用のデータとしては活用されてこなかった。その理由の一つとして、財務データ特有の外れ値や異常値、欠損値の存在などにより、データ分析を行うまでの初期整備に多くの労力を費やすこととなっていることが挙げられる。そこで、現在、CRDデータを学術研究用として扱いやすく標準化するための研究が行われている。CRDデータが標準化されることにより、初期データ整備が主眼ではない中小企業研究等で、CRDデータがこれまで以上に幅広く利用されることが期待される。また、この研究を通して、財務データ一般に関する標準的な整備手法が確立されることが期待される。今回は、CRDデータの標準化研究の中から、財務データに適した精度の良い欠損値補間方法について、サーベイ研究を紹介する。

【財務データの特徴】

企業の中で、株式市場に上場するような規模の企業はごく一部であり、多くが未上場の中小企業である。したがって、企業の財務データを集積すると、規模の多様性から、各財務項目の原数字の分布は、かなり偏った形状となる。その中でも、CRDデータは、中小企業法で定められている中小企業規準を満たしている企業の決算書データであるが、それでも分布の偏った性質は明確に残っている。（右図参照）



また、上場企業と異なり、厳密に決算書の作成基準が決められているわけではない中小企業の財務データに関しては、主要項目を除くと欠損値が比較的多くなる傾向があり、異常値なども散見される。CRDデータは、その提供元である全国の信用保証協会や金融機関のフィルターを通しており、さらにデータベースへの受け入れ時にクレンジングを行っているが、何らかの分析をする際には、外れ値や異常値、欠損値の影響を無視できない。そこで、分析を行う前にデータ整備を行う必要が生じるのであるが、本ポスターでは、データ整備の一つである欠損値補間に焦点を当て、財務データの特徴を考慮した補間方法について説明したい。

さて、非常に一般的に利用される欠損値補間方法として、平均値補間という方法が存在するが、その方法を記述したような分布の偏りがある場合に適用すると、多くの場合、補間値が真値から大きく乖離することは容易に想像できる。したがって、単純な平均値補間方法は、財務データに適用しても無意味であると考えられる。一方、企業規模や業績面については、時系列面での自己相関性が比較的強いと考えられることから、自己の前後期データの平均値補間については、ある程度予測力を持ちそうである。そこで、本研究では、前後期データ補間よりも精度の高い補間方法の開発を目標として設定したい。

なお、財務データに存在する欠損値の一部は、会計原則に従って、他の財務項目から逆算して補完することも可能になるが、本研究ではその方法が採用できない場合に、どのように精度良く、合理的に補間可能かをテーマとしている。また、欠損値に対しては欠損値を含むデータを分析用データから排除する（deletion）という方法も考えられるが、財務データの場合、欠損値が発生しているデータには、規模や業績の面で偏りが発生しているケースが多分にあるため、deletionは元のデータセットの特徴を保存する望ましい欠損値処理方法とは言い難く、その点も考慮可能な欠損値補間方法を研究することとしている。

【欠損値補間方法についての既存研究】

一般的な欠損値補間に関する既存の方法論としては、single imputation（単一値代入法）とmultiple imputation （多重代入法）という大きな流れがある。それらの方法論についての詳細は、次のように分類される。

分類	名称	内容・課題
single imputation	平均値補間法	平均値を挿入。標本分散が保存されない。
	ランダム補完法	フィールド毎に、一様分布から発生したランダムな値を補完する。標本分散は完全に保存されるが、他フィールドとの相関を考慮していないので、精度の高い補間は難しい。
	k-nn法 （k-nearest neighbor法）	k個の類似したレコードの値から補間値を作成。ある程度標本分散は保存されるとともに、比較的精度の高い補間が可能となる。また、類似性の計算方法や、補間値を作成する際の加重方法などの面で、応用事例が多い。
multiple imputation		複数の補間値を繰り返し計算から算出し、その平均を補間値とする。標本分散が比較的保存される。

既存研究をサーベイした限り、ファイナンス分野の論文では、imputationに関する論文はそれほど多くなく、特にsingle imputationに関してはほとんど見られない。

ファイナンス分野におけるmultiple imputationに関する研究では、欠損値補間をすることにより、財務データから企業の信用力を測る信用力判定モデルの精度を向上させることが可能かどうかを確認している。

single imputationに関する既存研究では、遺伝子マイクロアレイの時系列データや工数予測に関するクロスセクションデータに対して、k-nn法を適用もしくは応用した欠損値補間方法を、他の補間方法（平均値補間法等）と比較して、精度の良い方法として提案している。

上述の既存研究では、財務データの特徴である偏った分布である点、自己相関を有するデータである点などに着目した欠損値補間は行っていない。したがって、このような財務データの特徴にも既存研究が十分に適用可能かどうかを始めに確認し、それほど十分な精度が確保されないということであれば、既存の方法論を改善する必要がある。これまでの研究段階では、single imputationのうち、単純なk-nn法では、自己の前後期データ補完ほどの精度は見い出せなかったが、k-nn法を応用することにより、自己の前後期データ補完を上回る精度が実現できるようになっている。

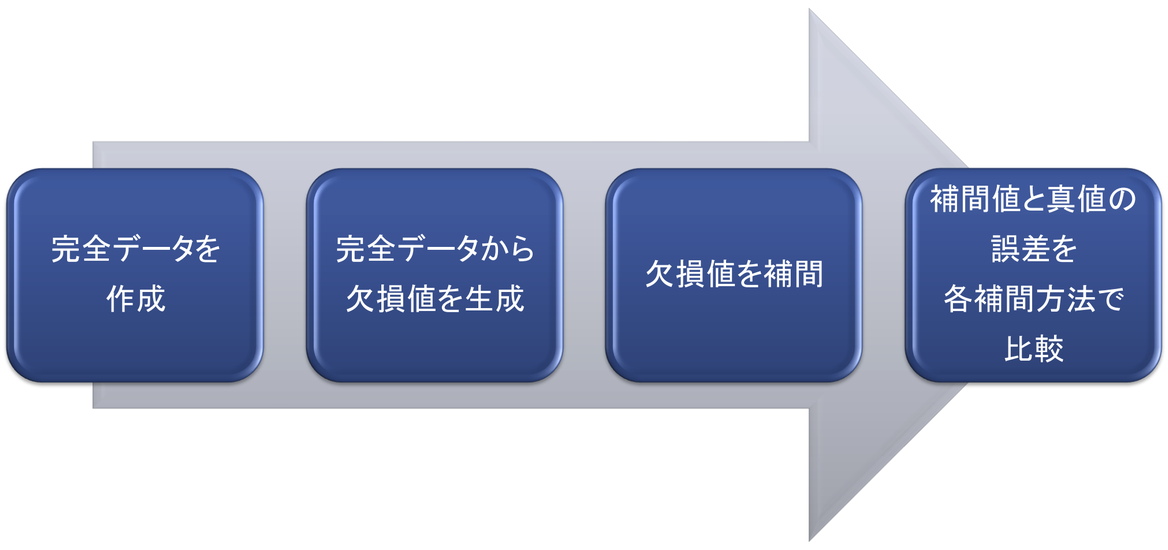
【欠損値補間の精度評価】

最後に、欠損値補間の精度に関する計測方法について、既存研究から紹介する。欠損値補間の精度評価方法については、大別すると以下の2種類の方法が存在する。

① モデル精度の向上を目的とするもの。
② 真値が既知のデータを用いた誤差評価によるもの。

ファイナンス分野では、欠損値補間を主目的とするのではなく、欠損値を補間して信用力判別モデルによる判定結果を出すことが最終目的となることが多く、その場合、①の方法が採用されることとなる。ただし、この場合、欠損値補間方法の良し悪しは、採用モデルによって異なるケースも考えうる。そこで、欠損値補間方法そのものの良し悪しを判断するためには、②の方法により、直接的に補間の精度を確認するのが良い。

ここで、②の方法についてやや詳しく説明しておく（下の手順図参照）。まず、欠損値を含まない完全データを用意する。次に、完全データから欠損値を何らかの方法（全項目ランダム、特定項目限定ランダム等）で発生させ、それに対して様々な欠損値補間方法でそれぞれ補間を行う。最後に、補間値と真値の誤差を、それぞれの補間方法で比較する。



上記最後における補間値と真値との誤差評価指標については、いくつかの指標が提案されているが、一般的と考えれられる以下の二つの指標を紹介する。

a. 規準化平均絶対誤差（NMAE：Normalized Mean Absolute Error）

$$NMAE = \frac{\sum_{i,j \in true} \left| \hat{z}_{ij} - z_{ij}^{true} \right|}{n}$$

ここで、z_{ij}^{true} ：完全データにおける真の値．$\hat{z}_{ij}$ ：補間値、n：欠損数。

b. 規準化平均平方誤差（NRMSE：Normalized Root Mean Squared Error）

$$NRMSE = \sqrt{\frac{\sum_{i,j \in true} \frac{\left(\hat{z}_{ij} - z_{ij}^{true}\right)^2}{\sigma_j^2}}{n}}$$

ここで、σ_j：項目の標準偏差。